ARMY RESEARCH LABORATORY

A Model for Predicting Intelligibility of Binaurally Perceived Speech

by Angélique A. Scharine, Paula P. Henry, Mohan D. Rao, and Jason T. Dreyer

ARL-TR-4075 **April 2007**

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

DESTRUCTION NOTICE—Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

ARL-TR-4075 **April 2007**

A Model for Predicting Intelligibility of Binaurally Perceived Speech

Angélique A. Scharine and Paula P. Henry
Human Research and Engineering Directorate, ARL

Mohan D. Rao and Jason T. Dreyer
Michigan Technological University

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2007	Final	October 2005 through September 2006

4. TITLE AND SUBTITLE

A Model for Predicting Intelligibility of Binaurally Perceived Speech

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Angélique A. Scharine and Paula P. Henry (both of ARL), Mohan D. Rao and Jason T. Dreyer (MTU)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, MD 21005-5425

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-4075

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Predicting and modeling intelligibility of monaurally or binaurally presented speech is difficult because it depends primarily on the accuracy and interdependency of frequency, time, and spatial information arriving at the listener. Despite these complex relationships, a new pragmatic model is suggested for speech mixed with broadband noise. A form of the logistic regression function is used to characterize human performance data. The regression of these signal properties onto empirical speech recognition performance data estimates the relationship of these properties to speech recognition. This concept is illustrated by the modeling of human performance on Central Institute for the Deaf W-22 speech items presented monaurally and binaurally in both reverberant and non-reverberant conditions at different signal-to-noise ratios. Although the implementation of the present model is limited to the data considered, it is expected that other data can be modeled after the procedure outlined in this report. The model described is the first step in developing an objective binaural measure for predicting speech perception in noisy environments.

15. SUBJECT TERMS

binaural; objective measures; speech intelligibility

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Angélique A. Scharine
Unclassified	Unclassified	Unclassified	SAR	35	19b. TELEPHONE NUMBER (*Include area code*) 410-278-5957

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

INTENTIONALLY LEFT BLANK

1. Background

1.1 Existing Objective Measures of Speech Intelligibility

There are a number of objective measures of speech intelligibility (SI). The current American national standard addressing SI (American National Standards Institute [ANSI] S3.5, 1997) refers to the two most common objective measures of SI: speech intelligibility index (SII) and speech transmission index (STI). Other objective measures of SI include percentage articulation loss of consonants (%ALcons)[1] and the direct-to-reverberant energy ratio.

The SII described in ANSI S3.5 (1997) is a revision of the articulation index (AI) found in the previous version (ANSI, 1969). AI was developed at Bell Labs in the late 1940s and is based on the proportion of the speech energy to the noise energy in a number of frequency bands. The number of bands and center frequencies of the bands differ in various implementations of the standard (e.g., French & Steinberg, 1947; Beranek, 1947; Fletcher & Galt, 1950; Kryter, 1952; Mueller & Killion, 1990; Pavlovic, 1991). The signal-to-noise ratio (SNR) in each frequency band is multiplied by an importance function for that band, based on the usefulness of information in this band to speech intelligibility, and the resulting index is a value between 0 and 1 representing the proportion of speech information that is audible. AI was the sole method described in the earlier standard, whereas it is one of two methods described in the SII. The SII differs from AI in that the weightings given to relative importance of various frequencies to speech intelligibility have been revised for improved accuracy. The SII also includes parameters to adjust for upward spread of masking and the standard speech spectrum level. In addition, the SII calculation details have been adapted to computer implementation rather than to manual chart-type computations.

STI was developed in the TNO[2] Laboratory in Holland by Steeneken and Houtgast (Steeneken & Houtgast, 1980; Houtgast, Steeneken, & Plomp, 1980) and is based on the SNR and impulse response of the transmission system. The STI model uses a test signal that has a speech-shaped spectrum and is modulated at a number of frequencies. At the receiving end of the communication system, noise, signal distortions, and reverberation in the system decrease to some extent the depth of each individual modulation frequency. Reductions in the modulation depth are associated with loss of intelligibility. We measure the changes in the depth of modulation by calculating a modulation transfer function (MTF) for each modulation frequency in each of a number of specific frequency bands. The resulting MTF values are converted into "equivalent speech-to-noise ratios" that are combined to form the STI, which is similar to the AI and can vary from 0 to 1 (Wijngaarden & Houtgast, 2004). SII incorporates STI for those situations in which SII is inappropriate. For example, STI is more appropriate for the measurement of SI in reverberant environments.

[1] This machine measure of intelligibility is computed from measurements of the direct-to-reverberant energy ratio and early decay time and is specified in percent.

[2] TNO = Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek.

Most objective measures predict SI quite accurately for the middle range of intelligibilities between 20% and 80% (Hargus & Gordon-Salant, 1995; Humes, Boney, & Loven, 1987; Steeneken & Houtgast, 2002). However, their usefulness is limited by a number of factors. First, the current objective measures of SI must be adjusted to account for the kind of speech material used. For example, SII (ANSI, 1997) has different importance functions for nonsensical syllables and monosyllabic words (NU-6, Northwestern University Auditory Test No. 6, Tillman & Carhart, 1966; CID W-22, Central Institute for the Deaf, Hirsh et al., 1952; DRT, Diagnostic Rhyme Test, Fair-2banks, 1958; and MRT, Modified Rhyme Test, House, Williams, Hecker, & Kryter, 1963) and short passages (CST, Connected Speech Test, Cox, Alexander, & Gilmore, 1987, and SPIN, Speech Perception in Noise, Kalikow, Stevens, & Elliott, 1977). There is an option to adjust SII, depending on the vocal effort of the talker. Some but not all objective measures account for the acoustic characteristics of the space in which speech occurs. For example, AI does not account well for moderate or severe reverberation in the environment. In the case of reverberant environments, the use of STI is more appropriate because reflected energy reduces the depth of modulation in the test signal and is thus accounted for by this index. The changes in AI, now called SII, improve its performance in reverberant environments by the incorporation of STI.

Note, too, that AI, STI, and SII are all indices and are not intended to directly predict SI. Instead, a transfer function is needed to convert the index scores to the predicted SI in percent correct. The transfer functions are included in the most recent ANSI standard for a number of common types of speech material.

1.2 How Binaural Measurement Can Improve Objective Tests

SI is affected by the location of the listener relative to the speech and the noise sources in the environment. When one is listening binaurally, changes in these relationships and the orientation of the human head affect SI. However, current objective measures of SI are based on measurements taken from a single microphone. These SI measures are essentially assuming monaural listening in the background of non-directional noise, and the resulting predictions can be considered the worst case scenario for SI. Although such worst case scenario data are sometimes very valuable, they are not realistic in terms of predicting the actual SI experienced by a human listener with two ears and may lead to incorrect and costly technical and operational decisions.

More specifically, the two measures described in ANSI (1997) require that both the target speech and the noise be co-located or omnidirectional. As a result, the actual SI data obtained will differ from the predicted data, depending on whether these requirements are met. Further, it would seem that these requirements are a special case, that noise in most real-world environments is not completely diffuse or co-located with the talker[3].

Considerable experimental evidence shows an advantage in speech perception for binaural over monaural listening (Drennan, Gatehouse, & Lever, 2003; Drullman & Bronkhorst, 2000; Gallun,

[3]In fact, one can predict that most talkers would be intelligent enough to move away from the noise source if it were possible!

Mason, & Kidd, 2005). The binaural advantage is thought to be attributable to several separate effects (Culling & Summerfield, 1995; Edmonds & Culling, 2006; Freyman, Balakrishnan, & Helfer, 2001; 2004). First, the listener can take advantage of the ear that is nearest to the target speech and farthest from the distracting noise and move his or her head appropriately. Second, the listener can take advantage of localization cues that allow him or her to spatially segregate the target speech from the noise. Further, the auditory system is able to correlate information from each ear to reduce masking from reverberation (Libbey & Rogers, 2004). These factors result in speech perception that is better than that predicted by monaural measurement methods and their associated indices, especially during the testing conditions specified in existing standards.

1.3 Purpose and Objective

Natural human listening involves two ears. Therefore, objective measures of SI based on measurement from a single microphone are not likely to predict SI accurately for various spatial conditions and SNRs. Conversely, they are likely to under-predict human performance in low SNR conditions, such as those experienced by the Soldier or other person in a noisy environment. An objective measure of speech intelligibility that uses binaural input and information about the surrounding space and thus more closely accounts for the capability of human hearing, would be more realistic, accurate, and useful in predicting SI performance in real-world environments.

In order to create an objective measure of SI, the information needed by the listener in order to accurately recognize speech must be determined. Speech is a time-varying complex sound resulting from changes in the shape and actions of the vocal mechanisms of the talker. In order to process speech, the listener must extract spectral (frequency content) and temporal (timing) information from the sound waveform. In noisy environments, the sound arriving at the listener's ears contains not only speech energy but also non-speech energy or noise that adversely affects the perception of speech. Because listeners must use selective information within a noisy signal in order to discern the speech from the noise, speech recognition can be treated as a problem requiring analytical listening.

Noise is an unwanted signal that masks speech signals and interferes with speech recognition. Masking of one sound by another has two forms: energetic and informational (Best et al., 2005). Energetic masking occurs when the energy contained in the competing sound signal masks the energy contained in speech. Energetic masking is thought to be the primary source of masking when speech occurs in the presence of a stationary background noise with characteristics similar to white or pink noise. The amount of energetic masking is primarily related to SNR and spectro-temporal properties of both speech and noise. Informational masking is additional masking that cannot be accounted for by energetic masking and is thought to be caused by cognitive and attentional aspects of noise. Masking noises that have a form of another speech signal, attractive/ emotional music, sound effects, and sporadic unexpected impulse noises are strong informational maskers.

Previous measures of SI have focused on the effects of energetic masking, in large part because they are based on models of the effect of the physical characteristics of noise on perception of spectro-temporal speech information. However, even in the presence of what is considered to be purely energetic maskers, SI varies significantly as a function of the speech material used, the talker reading the material, and the acoustic characteristics of the environment. The higher level cognitive processes that are thought to cause informational masking cannot be accounted for solely by physical measurements of the sound field. Modeling of informational masking requires several additional layers of information about the listeners beyond the information about their hearing ability. However, several elements of informational masking are highly dependent on binaural listening and therefore, any serious attempt to model the effects of informational masking on SI must consider binaural listening. This constitutes an additional challenge in developing an objective binaural measure of SI.

To a large extent, the effects of reverberation have also been disregarded in speech intelligibility metrics. Almost all real-world environments contain some degree of reverberation, and this affects speech recognition. The incorporation of reverberation into a speech intelligibility index is critical to accurately predicting human performance.

The objective of the current study was to model monaural and binaural speech recognition of the CID W-22 test items as a function of the physical characteristics of the signals arriving at two ears of the listener during one set of specific spatial conditions and two degrees of reverberation. To develop the model and to assess its goodness of fit, a set of human performance data was collected during the same environmental conditions as considered in the model. Thus, the study involved determining the performance intensity (PI) function for the W-22 speech test during specific test conditions, fitting the model to the test data, and assessing its goodness of fit. A very high goodness of fit for this limited set of data was considered the first step in the development of the future more flexible and robust model based on the same algorithm.

2. Method Used to Collect Human Performance Data for W-22 Items

Both modeling and validation of the new binaural SI measure addressed in this study required creation of special monaural and binaural W-22 sound files containing speech and noise to create various SNRs used in the study. Four SNRs and two reverberant (RT) conditions were used in the study. Together with monaural and binaural modes (MODE), these conditions resulted in 16 W-22 recordings.

2.1 Recording System

Figure 1 shows a diagram of the recording system used to create sound files for the study. A Knowles Electronic Manikin for Acoustic Research (KEMAR[4]) with two microphones situated at the position of the eardrums was placed in the center of a small sound-treated booth. A monophonic microphone was positioned just above the head of the KEMAR pointed toward the front. Two loudspeakers delivering speech and noise signals were placed 1 meter away from the center of the KEMAR head at ±45 degrees' azimuth. Technical specifications of the recording session follow.

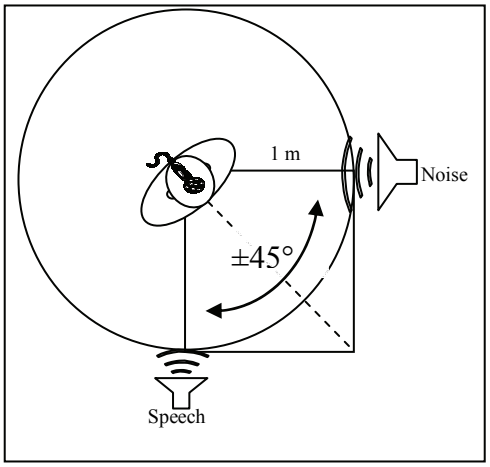

Figure 1. Recording system used to create sound files for the study.

2.2 Speech Material

The speech material was the "Auditec of Saint Louis" recording made by a male talker reading words from four CID W-22 lists (Hirsh et al., 1952). Each list consists of 50 phonemically balanced, monosyllabic English words. All recorded words had been leveled to produce the same peak root mean squared (rms) reading. These recordings were played back by a small Mission 780 loudspeaker (5-inch bass driver and 1-inch high range driver) with a frequency response rated to be 65 Hz to 20 kHz (±3 dB). This loudspeaker was chosen to mimic the directivity of speech and was placed 1 meter from the center of the KEMAR head at an angle of +45 degrees from the head median plane, as shown in figure 1. The output of the loudspeaker was calibrated with a 01dB Symphonie[5] system (Symphonie sound card and dbfa software) and a Microtech Gefell measurement microphone (sensitivity = 41.2 mV/Pa; frequency response = 3.5 Hz-20 kHz [± 2dB]). This same microphone was used as the monophonic microphone in the study (see figure 1). The calibration signal was a concatenated sample of several speech items spliced together. The

[4]Knowles Electronic Manikin for Acoustic Research is a trademark of Etymotic Research, Elk Grove Village, Illinois.

[5]01dB Symphonie and dbfa are trademarks of 01dB-metravib, Limonest Cedex, France.

reproduction system was adjusted to produce a speech signal at 70 dB sound pressure level (SPL) at the monophonic microphone location.

2.3 Background Noise

A 5-minute looped recording of pink[6] noise was played from a large PSB[7] Stratus loudspeaker (10-inch bass driver, 6-inch mid-range driver, and 1-inch tweeter). The on-axis frequency response is rated to be 31 Hz to 21 kHz (±3 dB). The loudspeaker was placed 1 meter from the center of the KEMAR head at an angle of -45 degrees from the head median plane, as shown in figure 1. The noise level at the loudspeaker output was calibrated with the same equipment as before and set at four intensity levels of 70, 73, 76, and 79 dB SPL in order to create the four SNRs of 0, -3, -6, and -9 dB used in the study.

2.4 Monaural and Binaural Test Recordings

2.4.1 Monaural Recording

The Microtech Gefell measurement microphone was placed approximately 2 inches above the KEMAR head, facing 0 degrees' azimuth. Its output was connected to the laptop computer via the 01dB Symphonie sound card, and the dbfa software was used to make the recordings.

2.4.2 Binaural Recording

Two Etymotic ER-11 microphones (sensitivity = 50 mV/Pa; frequency response = 250 Hz to 10 kHz [±2.5 dB]) were placed in the KEMAR ears to make the binaural recordings. The microphones were also connected to a laptop computer identical to the one used for the monaural recordings via a 01dB Symphonie sound card, and the dbfa software was used to make the recordings.

2.5 Reverberation

The binaural and monaural recordings at all SNRs were convolved with the impulse response of a large hall (the "church" pre-set in Adobe Audition[8] software: RT_{60} = 1.5 s) to create two reverberant conditions (RT = 0 s and RT = 1.5 s). RT = 0 s means that no reverberation was added to the recordings.

[6]Pink noise is random noise for which there is equal energy in all octaves. In terms of power at a constant bandwidth, pink noise falls off at 3 dB per octave.

[7]not an acronym

[8]Adobe and Audition are registered trademarks of Adobe Systems, Inc.

2.6 Human Performance Data

2.6.1 Participants

Twelve adults, aged 18 to 40 (mean = 29 years) with symmetrical normal hearing, volunteered to participate in the study. The participants were recruited from both the Government work force at the Aberdeen Proving Ground, Maryland, and the civilian population in surrounding communities. Each listener had pure-tone air conduction hearing thresholds ≥20 dB HL (hearing level) at octave frequencies from 250 through 8000 Hz (ANSI,) and no history of otologic pathology. The difference between pure-tone threshold HLs in both ears was no greater than 10 dB at any test frequency.

2.6.2 Experimental Task

Participants were presented with an individual word and instructed to use a computer interface to select the word heard from a list of 50 words. After selecting the word, the participant clicked on another computer button to initiate the presentation of the next word. All words for a particular list and condition were presented, with no repetitions in a single block.

2.6.3 Counterbalancing of Stimuli

There were 64 lists of stimuli (four lists of words × two MODEs × four SNRs × two RTs). Each listener heard one list of words in each of the 16 (4 SNRs × 2 MODEs × 2 RTs) conditions. A Latin square design was used to create a set of lists by the assignment of one of the four lists to each condition so that all four lists were used at each SNR level and so that no list was heard in a particular condition more than once (see table 1 for an example). Four such sets were made and each participant was assigned one of these sets so that each set was used for three participants. The lists in a set and items within each list were presented in a different random order for each participant.

Table 1. Sample counterbalanced design.

	Monaural		**Binaural**	
SNR (dB)	**RT=0**	**RT=1.5**	**RT=0**	**RT=1.5**
0	List 1	List 4	List 2	List 3
-3	List 2	List 1	List 3	List 4
-6	List 3	List 2	List 4	List 1
-9	List 4	List 3	List 1	List 2

2.6.4 Apparati

The words were saved as individual sound files and presented by computer via AKG[9]240F headphones. All sound files were inversely filtered to remove the frequency effects of the

[9]AKG is a trademark of AKG Acoustics GMBH, Vienna, Austria.

headphones before the items were presented to the listeners. A Symetrix[10] SX204 headphone amplifier was used to control the level of presentation. A Casella CEL[11] Ltd[12] 573.C1R sound level meter attached to a Brüel & Kjær (B&K) 4153 artificial ear[13] was used to calibrate the level of the 0 SNR recording of background noise (without speech) so that it measured 70 dB SPL when played from the headphones. For the binaural conditions, listeners were presented with the recordings made through the two ears of the KEMAR in the same setup as the KEMAR head had been during the recordings. For the monaural conditions, listeners were presented with the recordings made through the monophonic microphone above the head of the KEMAR, which was split to both ears (split monaural).

2.6.5 Test Data

The human performance data are shown in figure 2. For a given SNR, listener performances were much better for the binaural (dichotic) conditions than for the split monaural (diotic) conditions.

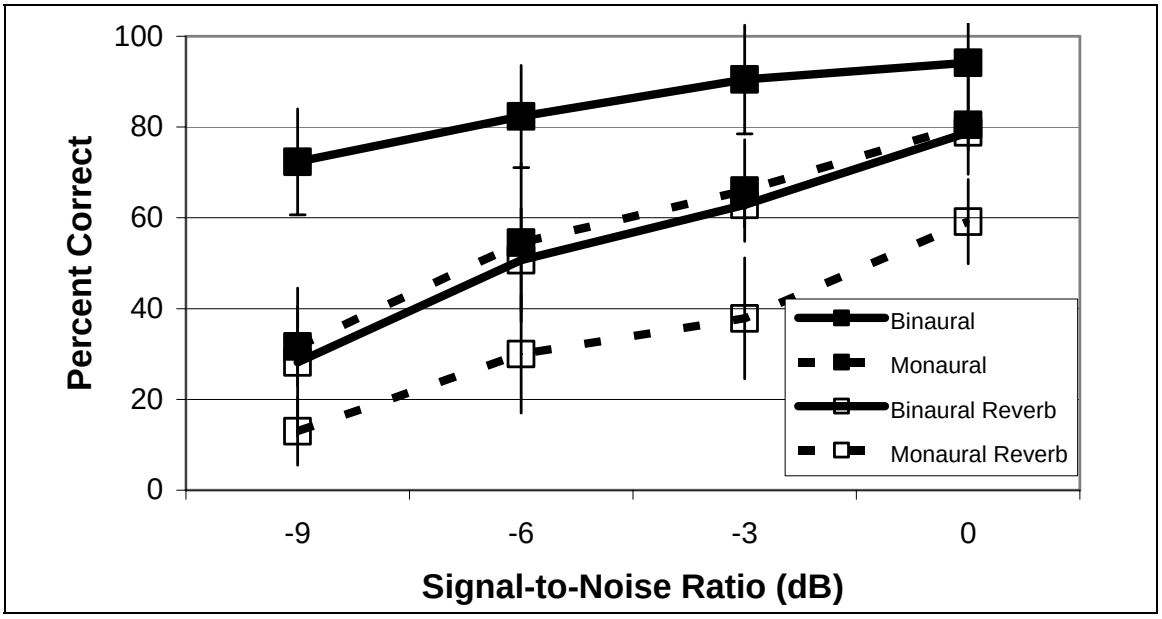

Figure 2. Percent correct obtained by human listeners for each of the four experimental conditions at each SNR.

[10]Symetrix is a trademark of Symetrix, Inc., of Mountlake Terrace, Washington.
[11]not an acronym
[12]Casella CEL Ltd is a trademark of Casella, Amherst, New Hampshire.
[13]Artificial Ear Type 4153 is a trademark of Brüel & Kjær Sound & Vibration Measurement A/S, Denmark.

3. Methods Used to Collect Human Performance Data on Callsign Acquisition Test (CAT) Items

In the initial stage of model development, before collection of W-22 data, the model algorithm was tested with previously collected data on speech recognition obtained with the CAT (Rao & Letowski, 2006). The purpose of using the CAT data was to get an early indication of how the algorithm "behaved" for different test conditions. Since there were significant differences in the data collection methods between the CAT and W-22 data, the following sections describe how the CAT data were obtained.

3.1 Recordings

The speech material was an in-house recording made by a male talker reading phrases of the CAT (Blue, Ntuen, & Letowski, 2004; Rao & Letowski, 2006). The CAT consists of 126 test items. A single test item (or call sign) consists of a word selected from a set of 18 two-syllable words taken from the military phonemic alphabet (Alpha-Zulu) and a number selected from a set of seven one-syllable digits (1 to 8 except 7) resulting in a three-syllable phrase, e.g., Bravo Five. The recorded words were equalized to produce the same peak rms values and filtered to remove the frequency shaping that occurs through the use of headphones.

Three background noises were used in collecting the CAT data: white noise, pink noise, and speech babble. The sound pressure level of the test items was set at 70 dB. The level of the background noise was then adjusted to create 3 SNR levels: -6, -9, and -12 dB. The speech and noise materials were presented diotically[14] over headphones. An IBM[15] personal computer and custom in-house software were used to control presentation order of the test items and to collect listeners' responses. The talker and the computer software were the same as later used in W-22 recordings and testing. More information about recordings and testing procedure used in the CAT study is presented in Rao and Letowski (2006).

3.2 Human Performance Data

3.2.1 Participants

A group of 18 listeners between the ages of 18 and 25 participated in the study. The participants were recruited from both the Government work force at the Aberdeen Proving Ground, Maryland, and the civilian population in surrounding communities. Hearing criteria were as described in section 2.6.1.

[14]That is, the same signal is presented to both ears. Normal listening is dichotic because the signal arriving at each ear is slightly different.

[15]International Business Machines

3.2.2 Apparati

The words were saved as individual sound files and presented by computer via AKG240F headphones. A Symetrix SX204 headphone amplifier was used to control the level of presentation. A 01dB Symphonie sound card and dbfa software attached to a B&K 4153 artificial ear were used to calibrate the level of the 0 SNR recording of background noise (without speech) so that it measured 70 dB SPL when played from the headphones.

3.2.3 Test Data

The human performance data obtained in the CAT study by Rao and Letowski (2006) are shown in figure 3. For a given SNR, white noise was found to be the easiest listening condition and pink noise was found to be the most difficult.

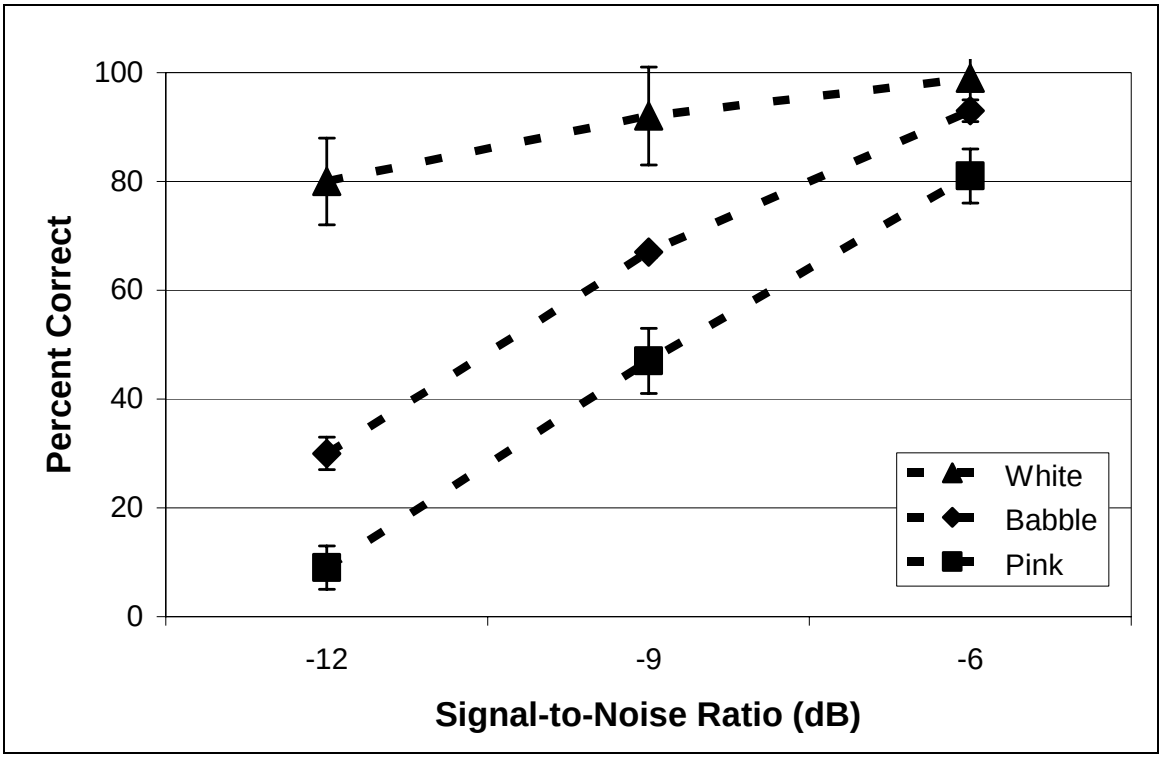

Figure 3. Percent correct obtained for the CAT test as a function of SNR for three different background noises.

4. Modeling Speech Intelligibility as Function of SNR

Speech intelligibility is generally defined as the percentage of test items recognized correctly (PC). When PC is obtained and graphed as a function of SNR, the resulting performance intensity (PI) function can be described as a sigmoid function having extremely high PC values at positive SNRs

and PC values close to 0% at negative SNRs. The standard logistic function (see appendix A) describing this curve is

$$PC(t) = \frac{1}{1 + e^{-at}} \tag{1}$$

The function $PC(t)$ describes a sigmoid that varies between 0 and 1 as a function of an independent variable, t. The parameter a determines the transition rate or the slope of the function. When t is equal to 0, the $PC(t) = 0.5$. In order to describe a PI function ranging from 0% to 100%, the right side of equation 1 is multiplied by 100. Further, the addition of a parameter a_0 allows the 50% value to be offset from t = 0. This can be expressed as

$$PC(t) = \frac{100}{1 + e^{a_0 + a_1 t}} \tag{2}$$

In the equation 2, a_0 determines the sigmoid's offset from 0 and a_1 determines its slope.

When a sigmoid function is used to model speech recognition data as a function of SNR, the SNRs are typically expressed in decibels which are base 10 logarithms of the actual SNR. In such cases, it is preferable to replace the natural logarithm base "e" in equation 2 by the decimal algorithm base "10". One such modified function is shown as equation 3. Equation 3 is equivalent to equation 2 except that the a_0 and a_1 in equation 3 are equal to the a_0 and a_1 in equation 2, divided by a constant.

$$PC(t) = \frac{100}{1 + 10^{-(a_0 + a_1 t)}} \tag{3}$$

As mentioned before, the SNR can be expressed as a logarithmic transformation of the ratio of the speech energy to the noise energy within a signal, written as

$$SNR = 10 \log_{10}\left(\frac{\sigma^2_{speech}}{\sigma^2_{noise}}\right) \tag{4}$$

in which σ^2_{speech} is the sample variance of the speech and σ^2_{noise} is the sample variance of the noise. In reality, people do not listen for the noise and speech separately but instead listen for the speech within the entire signal. The percentage of speech energy σ^2_{speech} within the entire signal energy $\sigma^2_{speech} + \sigma^2_{noise}$ can be written in terms of the SNR as

$$PC(SNR) = \frac{\sigma^2_{speech}}{\sigma^2_{speech} + \sigma^2_{noise}} \times 100\,. \tag{5}$$

If one solves for σ^2_{speech} in equation 4 and substitutes this into equation 5, the result can be rewritten as

$$PC(SNR) = \frac{100}{1+10^{-0.1SNR}} \quad (6)$$

Note that equation 6 has a form similar to equation 3 and has a value of 50% at an SNR of 0 dB. However, for different speech or testing conditions, the transition between 0% and 100% intelligibility may occur across different SNRs and at a different rate. Therefore, in order to express PC as a function of SNR, the above function must be modified to account for the shift in the 50% value along the SNR axis and for changes in the transition rate of speech perception. This modified function can be written to parallel function 3 and have a form

$$PC(SNR) = \frac{100}{1+10^{-(a_0+a_1SNR)}}, \quad (7)$$

in which the a_0 parameter is the offset value for PC = 50% in respect to SNR = 0 dB and the a_1 parameter is the transition rate (slope) of the curve at PC = 50%. This equation can then be used to model speech perception as a function of SNR.

In order to accurately model human speech recognition performance, it is desirable to have data points that describe the entire range of accuracy. However, it is quite difficult to obtain reliable performances at the end points and furthermore, realistic levels may require testing during conditions that are potentially dangerous to human participants. Thus, to fit a realistic approximation function to human data, the data need to be collected at a minimum of two points in the transition region between low (PC < 50%) and high (PC > 50%) intelligibility. These data points are needed to estimate the a_o and a_1 parameters of the intelligibility growth function.

The technique described was used to fit a function based on equation 7 to the CAT data. Figure 4 shows the resulting curves. Table 2 lists the best fit estimates of the model parameters. Inspection of figure 4 suggests that the slope of the three curves does not differ significantly and the changes in speech recognition performance attributable to background noise can be accounted for by changes in the offset parameter a_0. Table 3 shows the parameter values obtained for a_0 if a_1 is held constant at 0.25.

Table 2. Parameters obtained by fitting equation 7 to CAT speech recognition data.

Noise	a_0	a_1
Multi-talker Babble	2.4651	0.2388
Pink	2.3341	0.2728
White	3.4109	0.2381

Table 3. Parameters obtained by fitting equation 7 to CAT speech recognition data with fixed a_1.

Noise	a_0	a_1
Multi-talker Babble	2.58	0.25
Pink	2.13	0.25
White	3.56	0.25

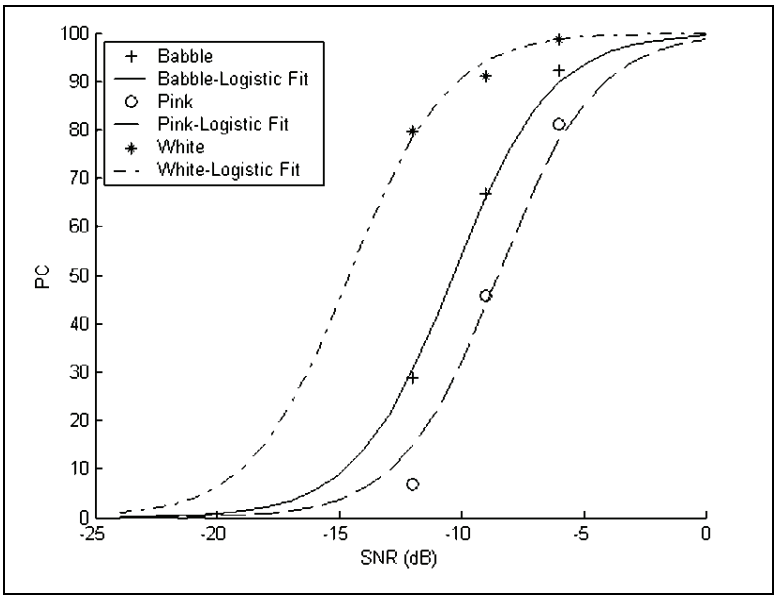

Figure 4. Percent correct obtained for the CAT test as a function of SNR for three different background noises. (Graphed lines represent the functions obtained by fitting equation 7 to the speech recognition data.)

Inspection of the speech recognition data obtained for the W-22 items reveals that the slope was the same for each of the four test conditions but differed from that obtained for the CAT data. Table 4 shows the estimated parameter values for a_0 when a_1 is held constant at 0.10. Once again, the differences in the testing conditions are accounted for by the offset a_0.

Note certain differences in tables 3 and 4 between the average a_1 estimated for each of the data sets. These differences are most likely attributable to the differences in speech material and some small high and low frequency differences in the spectral shape of the background noise resulting from the differences in the techniques used to equalize the headphone transfer functions in both studies.

Table 4. Logistic fit parameters of equation 7 for W-22 speech recognition data with fixed a_1.

Noise	a_0	a_1
Binaural	1.30	0.1
Monaural	0.62	0.1
Binaural - Reverb	0.55	0.1
Monaural - Reverb	0.14	0.1

5. Modeling Binaural Speech Intelligibility

Consideration of the estimated parameters illustrates the limitations of using SNR to predict SI. The parameters obtained will differ with every change in test condition (background noise, speech material used, or acoustic condition). For a given SNR, SI will also vary, depending on the

position of the listener relative to the target speech and the position of the target speech relative to the position of the noise. Therefore, in order to develop a useful objective metric of SI that is accurate during a number of conditions, it is necessary to consider binaural listening and to select some measurable aspect of the recorded waveform that varies as a function of these characteristics. To model binaural speech perception, we must be able to accurately predict the differences in SI for monaurally or binaurally presented speech during a variety of environmental conditions.

The regression equation 1 can be used in multiple forms to find the empirical relationship between any number of measurable aspects of speech and listener speech recognition including SNR, as shown in equations 2 through 7. The multiple regression form of the equation for $PC(SNR)$ is written as

$$EPC \equiv \frac{100}{1+10^{-\mathbf{Xa}}}, \tag{8}$$

in which $\mathbf{X}$ is the variable vector and $\mathbf{a}$ is the parameter vector.

One modeling effort attempted to fit the environmental data to successively higher order auto-regression (AR) equations (Rabiner & Schafer, 1978). AR separates the correlated components (speech) from the uncorrelated components (noise). As the order of the AR equation increases (for example, from m to n) and greater proportions of a signal are accounted for by the correlated portion, the residual sum of squares gets smaller. The change in the residual sum of squares for $AR(m)$ to $AR(n)$ is denoted as $R(n,m)$. SI was modeled as a function of this change by fitting equation 8 to the data, where $\mathbf{X} = [1 \quad 10\log_{10}(R(0,1)_N + 1) \quad 10\log_{10}(R(1,2)_N + 1) \quad \text{SNR}]$ and $\mathbf{a} = [a_0 \quad a_1 \quad a_2 \quad a_3]^T$ was the parameter vector.

This method worked well to distinguish between background noises that varied as a function of spectral content (see figure 5). However, it predicted no differences in performance because of the presence of reverberation, and most importantly, it did not fit binaural data well (see figure 6).

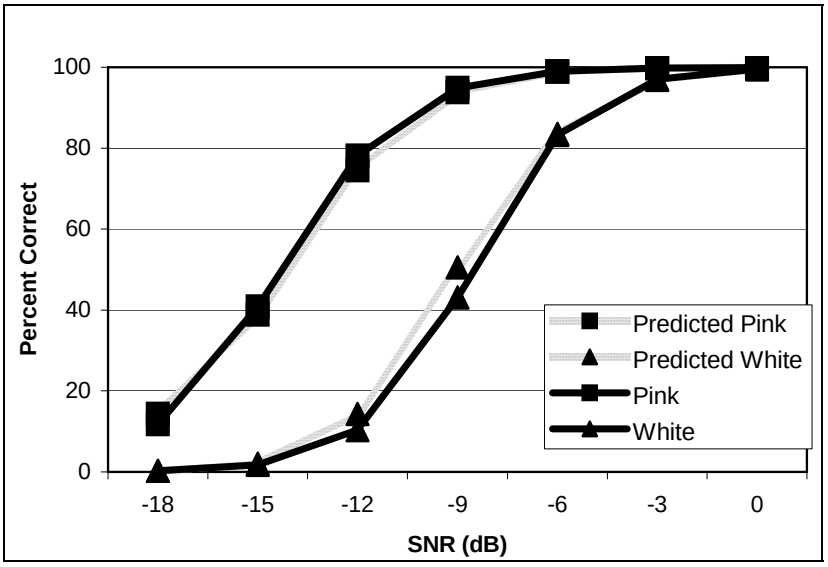

Figure 5. Predictions for monaural data in different steady state noise conditions for the first model as compared with human performance data.

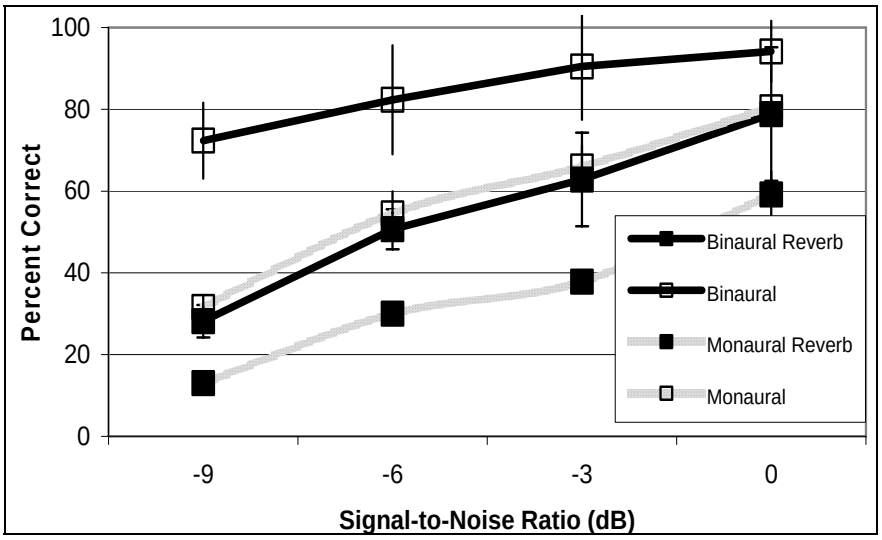

Figure 6. Predictions for binaural data from the second model as compared with human performance data.

A second modeling effort based on differences in low frequency sound energy was considered for fitting both the monaural and the binaural data. Once again, equation 8 was used as a starting point but with **X** defined as $\mathbf{X} = \begin{bmatrix} 1 & \max(M) & \min(M) & p & SNR \end{bmatrix}$ and **a** defined as $\mathbf{a} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \end{bmatrix}^{\mathrm{T}}$. The matrices **X** and **a** and the justifications for their use are described next.

5.1 "M"

Low frequency changes in the speech signal have been shown to disrupt speech recognition more than higher frequencies attributable to upward spread of masking. Although most speech energy is contained in the frequency range between 300 and 3000 Hz, it can be argued that signal periodicity (with frequencies below 20 Hz) defines the temporal pattern of speech and is very critical to speech recognition. A number of studies have shown that temporal information is more important for speech recognition than for spectral information (Drullman, Festen & Plomp, 1994a, 1994b). Drullman et al. found that periodic components around 4 to 6 Hz dominate speech recognition. Houtgast, Steeneken, and Plomp (1980) used low frequency amplitude modulation as a primary component in their measure of SI, the STI. The argument is that noise, by reducing modulation, masks speech. By measuring reduction in the depth of modulation for a number of frequencies between 0.4 Hz and 20 Hz, the degree to which speech recognition is impaired can be predicted. Therefore, it makes sense to consider low frequency signal energy in a model of SI. If a large proportion of the overall energy is from frequencies below a certain cut-off, then a low-pass filter will remove much of the energy that is attributable to speech. For a particular SNR, the ratio of low frequency energy to the overall energy will be large. However, if there is more high frequency energy relative to the overall energy, that ratio will be smaller, suggesting that a larger proportion of the overall energy is attributable to speech. Thus, SI will change as a function of this ratio for a particular SNR, based on the spectral content of the signal and the noise.

If low frequency energy from noise is detrimental to SI, then perhaps differences in this energy attributable to monaural and binaural listening are caused by the filtering effects of the head-related transfer function (HRTF).

Because reverberation attenuates higher frequencies and increases lower frequency energy, this ratio should reflect the effects of reverberation by increasing as reverberation increases. In this way, a model based on the proportion of low frequency energy may be able to account for changes in SI attributable to reverberation.

In order to account for the different effects of low frequency energy on intelligibility of monaurally and binaurally perceived speech, we modeled the low frequency effect of the HRTF as a second order Butterworth low-pass filter with a cut-off frequency of 100 Hz. The frequency value of 100 Hz was experimentally derived; it was the cut-off frequency that allowed for the best prediction of the data from the sound files. Subsequently, the variance of the total signal, σ_{TOT}^2, was compared to the variance of the filtered signal, σ_{LP}^2. The resultant calculation of the relative energy of the LP signal is

$$M = 10\log_{10}\left(\frac{\sigma_{LP}^2}{\sigma_{TOT}^2}\right) \tag{9}$$

We wish to clarify here that *M* is in no way related to the *m* referred to in the STI measure. *M* in this model can be described as the proportion of low frequency energy relative to the entire signal for a particular channel (ear). The *m* used in STI is the MTF and is equivalent to the SNR described in equation 6 except that SNR is described as a proportion rather than a percent.

M is calculated for each channel (ear) independently. Min(*M*) and max(*M*) refer to the smaller and larger of these two values. By calculating *M* for each channel, the model may be able to capitalize on differences between the two ears for binaural signals. If one ear has a better ear advantage (more favorable SNR), thus increasing *M* for that ear, the model can use this information to predict improved SI. In the case of monaural or diotic data, the *M* value is the same for both channels.

5.2 Number of Channels (p)

A variable *p* (*p* = 1 or 2) is needed to inform the model if there are one or two channels of information.

5.3 SNR

Although monaural SNR, or SNR defined as the average SNR across both ears for binaural listening, is insufficient to predict SI, there is a clear relationship between SNR and SI for a given talker, talker effort, mode, acoustic condition, and spatial configuration. Therefore, it is included in the model.

With the speech recognition data obtained for the W-22 speech items, the parameter vector **a** for the W-22 test data was found to be $\mathbf{a} = [-0.5250 \quad -0.0398 \quad 0.2088 \quad 1.8481 \quad 0.0864]^T$. The predicted PCs described by this model are graphed with the actual performance data in figure 7. The R-squared value for the regression was 0.99, which is apparent from the close fit of the curves on the experimental data. This figure demonstrates that listener responses to monaural and binaural stimuli presented in different reverberant conditions can be predicted from a single model.

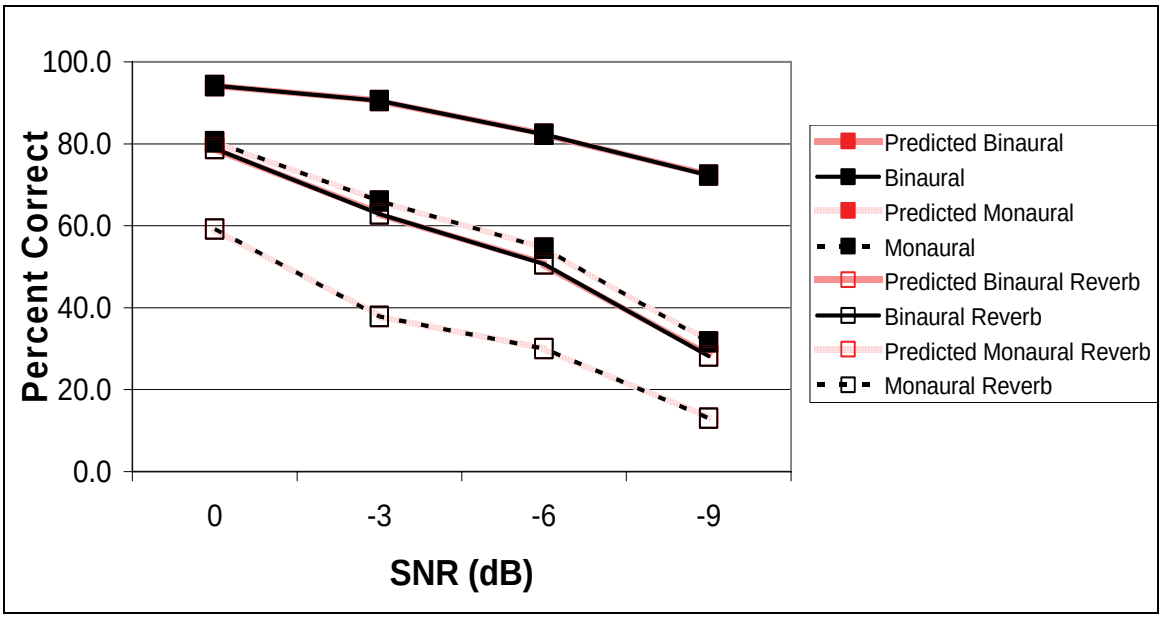

Figure 7. Estimated PC graphed with corresponding human performance data.

Equation 8 can be fit to any speech recognition data set in order to obtain parameters that can be used to predict SI performance during the same conditions as the data collected. It can also be used to predict SI (for the same speech material and the same testing environment) at SNRs other than the ones for which data have been obtained.

6. Discussion

SI, as found for the single set of data used in this study, was accurately modeled as a function of the selected variables: *M*, SNR, and the number of channels (*p*). This model was able to account for the differences in SI because of the change in the reverberation time from 0 to 1.5 s. It was also able to differentially model SI performance for binaurally and monaurally presented stimuli during the conditions tested in this study. Thus, the obtained data may be considered as meeting the requirements of a "proof of concept"; however, much work is still needed to develop a more general model.

The predictive value of the present model is currently limited to the speech material and testing conditions used in this study, which limits the generalizability of the model. It may be possible to "train" the model to other data sets in order to broaden its predictive value; however, this method would require a different set of parameters for each test condition. This may be desirable or necessary in some cases. For example, it is unlikely that a model based on recordings will accurately predict performance differences attributable to differences in speech material used. A different set of parameters may be needed to predict performance on each of the following: phonemes, syllables, words, sentences, short paragraphs, etc. However, for differences in SI that are attributable to spatial features such as the position of the talker relative to the listener and noise source and the directionality of the noise source, it is preferable to obtain a variable(s) that describe how SI changes as a function of spatial layout. It is currently unknown whether the *M* values can adequately account for changes in SI because of differences in spatial layout of the listening environment. Similarly, the variables used in a model need to account for changes attributable to acoustics, such as changes in room size or reverberation time. It is unknown whether *M* will continue to predict this for a large variety of spatial layouts. More investigation is needed in order to verify that appropriate binaural information in the sound signal is incorporated into the model.

This yet-to-be-determined variable may be in addition to *M* or it may replace it completely. It is unclear whether *M* provides any useful information that is generalizable to other contexts because it is unclear why *M* changes as a function of whether hearing is binaural versus monaural. Although low frequency sound energy should change as a function of the angle of entry into the ear canal because of reflectivity of the neck and torso and shadowing effects, the changes at higher frequencies are much more significant for SI. However, these changes should still be quite small compared to the overall variability in spectral energy.

One could argue that the prediction of improved SI in the binaural recording conditions is solely attributable to the differences between the locations of the microphones during the recordings. Recall that for the monophonic recordings, a single microphone was positioned above the KEMAR oriented toward the front (0 degrees). The binaural recordings were obtained through microphones at the position of the listener's eardrums. Therefore, the binaural recordings were shaped by the HRTFs whereas the monaural recordings were not. The ability to differentiate between monaural and binaural signals through low-pass filtering at 100 Hz would only hold true for this recording paradigm. Had the participants listened through each channel (ear) of the binaural recording separately plus through the true binaural recording, more realistic comparisons between the two conditions might have been made, but this may have also forced changes in how the model was created to differentiate between the two types of signals. Further data collection and analyses will need to be conducted in order to refute or verify this assertion.

Furthermore, it will be necessary to ascertain whether the differences in low frequency energy are not entirely attributable to technical factors that affected the recordings used for the collection of human speech recognition data. For example, the frequency range of the monaural microphone was approximately 3.5 Hz to 20 kHz and the frequency range of the microphones used in the

KEMAR is approximately 250 Hz to 10 kHz. Therefore, the amount of low frequency energy recorded by the KEMAR is necessarily much smaller. This has the effect of filtering most of the low frequency energy before the low-pass filter is used and *M* is artificially lowered. It is unlikely that the low-pass filter was capturing loss of SI attributable to loss of low frequency information. Despite these serious limitations of *M*, it was possible to model SI performance in different reverberation conditions. It is likely that the predictive power of the model for the differences in SI for monaural and binaural stimuli is attributable to the *p* variable. This limits the potential for the use of *M* to predict SI for a variety of spatial layouts.

Since binaural facilitation for SI is probably attributable to the use of spatial cues to segregate speech from noise, the ideal alternative to *M* is to use a variable that captures these binaural cues. Specifically, such a variable would need to incorporate phase and level differences between the ears. Bharitkar and Kyriakakis (2006) describe a method of creating three-dimensional (3-D) sound by creating a transfer function that reproduces the sound that occurs in one ear by filtering the sound that occurs in the other ear. Although our purpose is not to create 3-D sound, this filter contains the binaural information we need. Because the filter described is minimum phase, all phase information is preserved. If the phase information for the low frequency portion (100 to 1500 Hz) of the signal could be extracted, it could serve as a variable for use in the regression equation describe in equation 8. As long as sufficient low frequency energy is present in the speech signal (speech is usually in the range of 300 to 3000 Hz), the phase information present in the filter would be sufficient to spatially separate speech from a diffuse background noise.

Although binaural cues are insufficient to precisely localize a sound in azimuth, they are sufficient to localize a sound horizontally in the front (or rear) hemisphere. This would allow us to predict variations in SI as a function of the spatial relationship because it is unlikely that SI is different in the front and back. Some additional precision might be achieved if we look at the level differences in the frequency range between 1 and 3 kHz. Figure 8 shows the interaural level differences for a sound source at three azimuthal positions. However, since these differences are in the range of 0 to 10 dB, it may be a bit more difficult to detect these differences reliably when speech is combined with a noisy background. Some investigation is necessary to determine whether it can be used as a variable in the model.

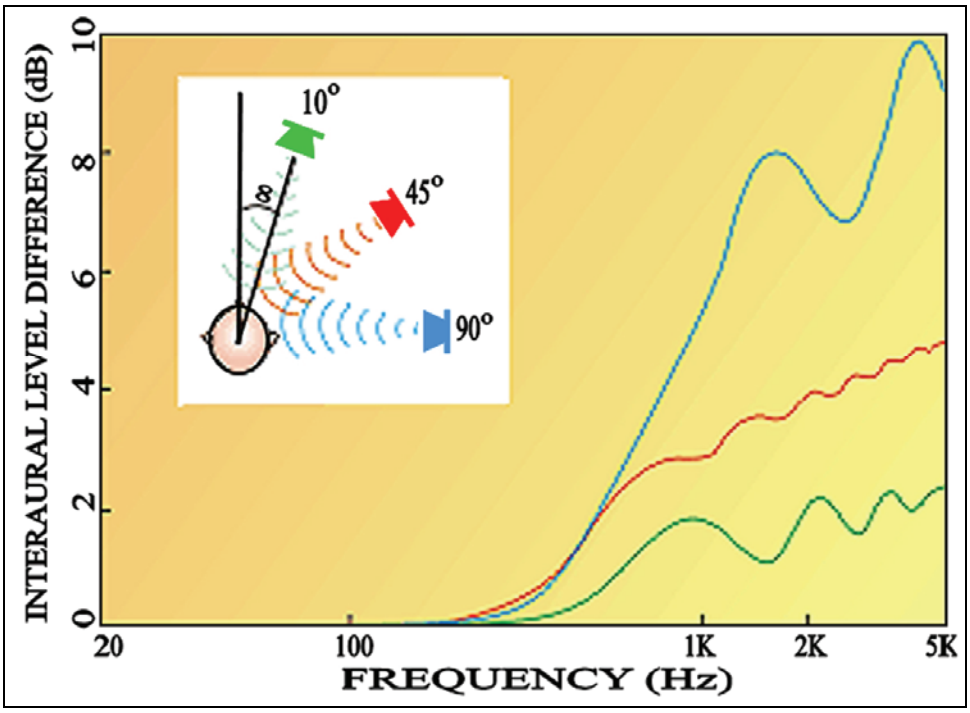

Figure 8. Interaural level differences, calculated for a source in the azimuthal plane defined by the two ears and the nose. (The source radiates frequency *f* and is located at an azimuth q of 10 degrees [green curve], 45 degrees [red], or 90 degrees [blue] with respect to the listener's forward direction. The calculations assume that the ears are at opposite poles of a rigid sphere [Hartmann, 1999].)

7. Conclusions and Recommendations

A regression equation with binaural summation was used to describe speech recognition performance based on human performance data. This relationship provides a link between human performance and several signal properties, namely, low frequency content, number of channels, reverberation, and SNR.

Currently, the model has only been fit to one set of human speech recognition performance data, and the parameter estimates depend on the specific characteristics of these data. These parameter estimates require validation by testing on other data collected during similar circumstances as well as investigations to eliminate technical factors that could have influenced the results.

Other environmental factors should be investigated in order to create a more universal model and allow for the development of a binaural speech metric based on speech recordings made in background noise. For example, the spatial location of the speech source with respect to the noise source, the amplitude distribution function of noise, and the rate of speech may also contribute to speech recognition scores.

The present work was also naturally limited by the use of only one type of background noise. Further work in model development should include fitting data sets (obtained during similar testing conditions) of monaural and binaural human performance on speech presented with different background noises and with different amounts of reverberation.

8. References

American National Standards Institute. American National Standard Methods for the Calculation of the Articulation Index; ANSI S3.5, New York, 1969, revised 1986.

American National Standards Institute. American National Standard Methods for Calculation of the Speech Intelligibility Index; ANSI S3.5, New York, 1997, revised 2002.

American National Standards Institute. American National Standard Specification for Audiometers; ANSI S3.6, New York, 2004.

Beranek, L. L. The Design of Speech Communication Systems. *Proceedings of the Institute of Radio Engineers* **September 1947**, *35* (9); 880-890.

Best, V.; Ozmeral, E.; Gallun, F. J.; Sen, K.; Shinn-Cunningham, B. G. Spatial Unmasking of Birdsong in Human Listeners: Energetic and Informational Factors. *Journal of the Acoustical Society of America* **December 2005**, *118* (6), 3766-3773.

Bharitkar, S.; Kyriakakis, C. *Immersive Audio Signal Processing*; Springer Science+Business Media: New York, NY, 2006.

Blue, M.; Ntuen, C.A.; Letowski, T.R. Speech Intelligibility of Callsign Acquisition Test (CAT) in a Quiet Environment. *Journal of Occupational Safety and Ergonomics* **2004**, *20*, 179-189.

Cox, R. M.; Alexander, G. C.; Gilmore, C. A. Development of the Connected Speech Test (CST). *Ear and Hearing* **1987**, *8* (Suppl.), 119S–126S.

Culling, J. F.; Summerfield, Q. The Role of Frequency Modulation in the Perceptual Segregation of Concurrent Vowels. *Journal of the Acoustical Society of America* **August 1995a**, *98* (2 Pt. 1), 837-846.

Culling. J. F.; Summerfield, Q. Perceptual Separation of Concurrent Speech Sounds: Absence of Across-Frequency Grouping by Common Interaural Delay. *Journal of the Acoustical Society of America* **August 1995b**, *98* (2 Pt 1), 785-797.

Drennan, W. R.; Gatehouse, S.; Lever, C. Perceptual Segregation of Competing Speech Sounds: The Role of Spatial Location. *Journal of the Acoustical Society of America* **October 2003**, *114* (4 Pt 1), 2178-2189.

Drullman, R.; Bronkhorst, A. W. Multichannel Speech Intelligibility and Talker Recognition Using Monaural, Binaural, and Three-Dimensional Auditory Presentation. *Journal of the Acoustical Society of America* **April 2000**, *107* (4), 2224-2235.

Drullman, R.; Festen, J. M.; Plomp, R. Effect of Reducing Slow Temporal Modulations on Speech Reception. *Journal of the Acoustical Society of America* **February 1994a**, *95* (2), 2670-2680.

Drullman, R.; Festen, J. M.; Plomp, R. Effect of Temporal Envelope Smearing on Speech Reception. *Journal of the Acoustical Society of America* **May 1994b**, *95* (5 Pt 1), 1053-1064.

Edmonds, B. A.; Culling, J. F. The Spatial Unmasking of Speech: Evidence for Better-Ear Listening. *Journal of the Acoustical Society of America* **September 2006**, *120* (3), 1539-1545.

Fairbanks, G. Test of Phonemic Differentiation: the Rhyme Test. *Journal of the Acoustical Society of America* **1958**, *30*, 596-600.

Fletcher, H.; Galt, R. H. The Perception of Speech and Its Relation to Telephony. *Journal of the Acoustical Society of America* **March 1950**, *22* (2), 89-151.

French, N.; Steinberg, J. Factors Governing the Intelligibility of Speech Sounds. *Journal of the Acoustical Society of America* **January 1947**, *19* (1), 90-119.

Freyman, R. L.; Balakrishnan, U.; Helfer, K. S. Spatial Release From Informational Masking in Speech Recognition. *Journal of the Acoustical Society of America* **May 2001**, *109* (5 Pt 1), 2112-22.

Freyman, R. L.; Balakrishnan, U.; Helfer, K. S. Effect of Number of Masking Talkers and Auditory Priming on Informational Masking in Speech Recognition. *Journal of the Acoustical Society of America* **May 2004**, *115* (5 Pt 1), 2246-2256.

Gallun, F. J.; Mason, C. R.; Kidd, G. Binaural Release From Informational Masking in a Speech Identification Task. *Journal of the Acoustical Society of America* **September 2005**, *118* (3), 1614-1625.

Hargus, S. F.; Gordon-Salant, S. Accuracy of Speech Intelligibility Index Predictions for Noise-Masked Young Listeners With Normal Hearing and for Elderly Listeners With Hearing Impairment. *Journal of Speech and Hearing Research* **February 1995**, *38* (1), 234-243.

Hartmann, W. M. How We Localize Sound. *Physics Today*, **November 1999**, 24-29.

Hirsh, I.; Davis, H.; Silverman, S. R.; Reynolds, E. G.; Eldert, E.; Benson, R. W. Development of Materials for Speech Audiometry. *Journal of Speech and Hearing Disorders* **September 1952**, *17* (3), 326-337.

House, A. S.; Williams, C.; Hecker, M. H.; Kryter, K. D. *Psychoacoustic Speech Tests: A Modified Rhyme Test*; ESD-TDR-63-403; U.S. Air Force, System Command Electronic Systems Division, **1963** *86*, 1-44.

Houtgast, T.; Steeneken, H. J. M.; Plomp. R. Predicting Speech Intelligibility in Rooms From the Modulation Transfer Function. I. General Room Acoustics. *Acustica* **1980**, *46* (1), 60-72.

Humes, L. E.; Boney, S.; Loven, F. Further Validation of the Speech Transmission Index (STI). *Journal of Speech and Hearing Research* **September, 1987**, 30 (3), 403-410.

Kalikow, D. N.; Stevens, K. N.; Elliott, L. L. Development of a Test of Speech Intelligibility in Noise Using Sentence Materials With Controlled Word Predictability. *Journal of the Acoustical Society of America* **1977**, *61*, 1337–51.

Kryter, K. D. Methods for the Calculation and Use of the Articulation Index. *Journal of the Acoustical Society of America* **November 1962**, *34* (11), 1689-1697.

Libbey, B.; Rogers, P. H. The Effect of Overlap-Masking on Binaural Reverberant Word Intelligibility. *Journal of the Acoustical Society of America* **November 2004**, *116* (5), 3141-51.

Mueller, H. G.; Killion, M. C. An Easy Method for Calculating the Articulation Index. *Hearing Journal* **September 1990**, *43* (9), 14-17.

Pavlovic, C. Speech Recognition and Five Articulation Indexes. *Hearing Instruments* **March 1991**, *42* (9), 20-23.

Rabiner, L. R.; R. W. Schafer. *Digital Processing of Speech Signals*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1978.

Rao, M. D.; Letowski, T. R. Callsign Acquisition Test (CAT): Speech Intelligibility in Noise. *Ear and Hearing* **April 2006**, *27* (2), 120-128.

Steeneken, H. J. M.; Houtgast, T. A Physical Method for Measuring Speech Transmission Quality. *Journal of the Acoustical Society of America* **1980**, *67*, 318-326.

Tillman, T. W.; Carhart, R. *An Expanded Test for Speech Discrimination Utilizing cnc Monosyllabic Words*; Northwestern University Auditory Test No. 6, Technical Report No. SAM-TR-66-55; USAF School of Aerospace Medicine: Brooks Air Force Base, TX, 1966.

Wijngaarden, S. van; Houtgast, T. Effect of Talker and Speaking Style on the Speech Transmission Index (STI). *Journal of the Acoustical Society of America* **2004**, *115*, 38-41.

Appendix A. Logistic Function Parameter Estimation

The parameters of the logistic functions can be estimated by various logistic regression techniques. Logistic regression can be expressed as a linear regression of transformed response variables. Two common transformations of response variables are logit and probit transformations. However, if the variables actually include 1 or 0, the logit transformation is undefined. Using nonlinear least squares (NLS) estimation techniques does not require transformation of the data and thereby allows estimation of the parameters without concern for extreme data values (inclusive of 0 or 1). Curves with NLS-generated parameters also often fit the data closer than those using logit or probit transformation, thus making its use preferable here.

The parameter estimation in this report used the following procedure used to fit the ARL listener data to the logistic function to the W-22 speech recordings.

The listener results were recorded as averages for the different test conditions, defined as the response vector **Y**. The average signal properties during these test conditions were then calculated and recorded in the prediction vector **X**. The columns of this vector were defined as the following measured parameters: constant, max(*M*), min(*M*), number of channels, and SNR. The subscripts on the average signal properties *M* were for the different test conditions (1 = mono, 2 = mono + reverb, 3 = stereo, 4 = stereo + reverb).

$$\mathbf{Y} = \begin{bmatrix} 80 \\ \vdots \\ 32 \\ 58 \\ \vdots \\ 14 \\ 92 \\ \vdots \\ 74 \\ 80 \\ \vdots \\ 31 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & \max(M_1) & \min(M_1) & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \max(M_1) & \min(M_1) & 1 & -9 \\ 1 & \max(M_2) & \min(M_2) & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \max(M_2) & \min(M_2) & 1 & -9 \\ 1 & \max(M_3) & \min(M_3) & 2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \max(M_3) & \min(M_3) & 2 & -9 \\ 1 & \max(M_4) & \min(M_4) & 2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \max(M_4) & \min(M_4) & 2 & -9 \end{bmatrix} \tag{A-1}$$

A multidimensional unconstrained nonlinear minimization routine was used to estimate the parameter vector $\mathbf{a}$ in the regression formulation (FMINSEARCH using the Nelder-Mead simplex direct search method in MATLAB[16]). The objective function was the sum of squares of the error (SSE) between the predicted value and the actual values of the response variable.

[16]MATLAB is registered trademark of the MathWorks.

$$SSE = \sum \left(\mathbf{Y} - \frac{100}{1 + 10^{-\mathbf{Xa}}} \right)^2 \qquad \text{(A-2)}$$

The initial values of the parameters, **a,** were considered zero for this NLS routine. The R-squared value was calculated by the following formulation:

$$R^2 = \frac{\sum \left(\frac{100}{1 + 10^{-\mathbf{Xa}}} \right)^2}{\sum (\mathbf{Y})^2} \qquad \text{(A-3)}$$

NO. OF COPIES ORGANIZATION

1 (PDF ONLY) DEFENSE TECHNICAL INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 US ARMY RSRCH DEV & ENGRG CMD
SYSTEMS OF SYSTEMS INTEGRATION
AMSRD SS T
6000 6TH ST STE 100
FORT BELVOIR VA 22060-5608

1 DIRECTOR
US ARMY RESEARCH LAB
IMNE ALC IMS
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
AMSRD ARL CI OK TL
2800 POWDER MILL RD
ADELPHI MD 20783-1197

2 DIRECTOR
US ARMY RESEARCH LAB
AMSRD ARL CS OK T
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR M DR M STRUB
6359 WALKER LANE SUITE 100
ALEXANDRIA VA 22310

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR ML J MARTIN
MYER CENTER RM 2D311
FT MONMOUTH NJ 07703-5601

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MZ A DAVISON
199 E 4TH ST STE C TECH PARK BLDG 2
FT LEONARD WOOD MO 65473-1949

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MD T COOK
BLDG 5400 RM C242
REDSTONE ARSENAL AL 35898-7290

NO. OF COPIES ORGANIZATION

1 COMMANDANT USAADASCH
ATTN ATSA CD
ATTN AMSRD ARL HR ME MS A MARES
5800 CARTER RD
FT BLISS TX 79916-3802

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MO J MINNINGER
BLDG 5400 RM C242
REDSTONE ARSENAL AL 35898-7290

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MM DR V RICE
BLDG 4011 RM 217
1750 GREELEY RD
FT SAM HOUSTON TX 78234-5094

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MG R SPINE
BUILDING 333
PICATINNY ARSENAL NJ 07806-5000

1 ARL HRED ARMC FLD ELMT
ATTN AMSRD ARL HR MH C BURNS
BLDG 1467B ROOM 336
THIRD AVENUE
FT KNOX KY 40121

1 ARMY RSCH LABORATORY - HRED
AVNC FIELD ELEMENT
ATTN AMSRD ARL HR MJ D DURBIN
BLDG 4506 (DCD) RM 107
FT RUCKER AL 36362-5000

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MK MR J REINHART
10125 KINGMAN RD
FT BELVOIR VA 22060-5828

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MV HQ USAOTC
S MIDDLEBROOKS
91012 STATION AVE ROOM 111
FT HOOD TX 76544-5073

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MY M BARNES
2520 HEALY AVE STE 1172 BLDG 51005
FT HUACHUCA AZ 85613-7069

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MP D UNGVARSKY
BATTLE CMD BATTLE LAB
415 SHERMAN AVE UNIT 3
FT LEAVENWORTH KS 66027-2326

NO. OF COPIES ORGANIZATION

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MJK J HANSBERGER
JFCOM JOINT EXPERIMENTATION J9
JOINT FUTURES LAB
115 LAKEVIEW PARKWAY SUITE B
SUFFOLK VA 23435

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MQ M R FLETCHER
US ARMY SBCCOM NATICK SOLDIER CTR
AMSRD NSC SS E BLDG 3 RM 341
NATICK MA 01760-5020

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MY DR J CHEN
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MS MR C MANASCO
SIGNAL TOWERS 118 MORAN HALL
FORT GORDON GA 30905-5233

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MU M SINGAPORE
6501 E 11 MILE RD MAIL STOP 284
BLDG 200A 2ND FL RM 2104
WARREN MI 48397-5000

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MF MR C HERNANDEZ
BLDG 3040 RM 220
FORT SILL OK 73503-5600

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MW E REDDEN
BLDG 4 ROOM 332
FT BENNING GA 31905-5400

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MN R SPENCER
DCSFDI HF
HQ USASOC BLDG E2929
FORT BRAGG NC 28310-5000

1 ARMY G1
ATTN DAPE MR B KNAPP
300 ARMY PENTAGON ROOM 2C489
WASHINGTON DC 20310-0300

NO. OF COPIES ORGANIZATION

ABERDEEN PROVING GROUND

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK (TECH LIB)
BLDG 4600

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK S FOPPIANO
BLDG 459

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR MR F PARAGALLO
BLDG 459

20 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR SD A SCHARINE
BLDG 520 APG AA